Can You Be More Explicit?
A Task and Dataset on Explicitations of Implicit Meaning

Laura Zeidler and Michael Roth
Natural Language Understanding Lab
University of Technology Nuremberg
{firstname.lastname}@utn.de

Abstract

Making texts clear and comprehensible has become an increasingly important topic in NLP. A possible strategy to enhance text comprehension is to make implicitly conveyed meaning explicit. To explore the role of explicit vs. implied meaning, we study cases of so-called explicitations, i.e. revisions of text in which implicitly conveyed content is made explicit. Using revision histories from wikiHow, we propose a rule-based approach to extract candidate explicitations and curate a human-annotated dataset in which explicitations are distinguished from insertions of new information. Our analyses show that while the extraction method is effective in retrieving relevant cases, distinguishing explicitations from new information is a challenging and often subjective task, reflecting differences in background knowledge and reasoning. Experimentally, we find off-the-shelf LLMs to achieve promising performance, with inconsistent gains from few-shot prompting and fine-tuning. In contrast, fine-tuned NLI models benefit consistently from supervised training and show stronger robustness under distribution shift. In sum, our findings show that the task is challenging, but also indicate that our annotated dataset contains informative signals that models can learn from, paving the way for further research on explicitations.

1 Introduction

Understanding the meaning of a text requires us to understand both what is explicitly written as well as to infer those aspects of meaning that are only implicitly understood (Carston, 2002). For instance, an instructional text may tell us to perform a particular step, but without specifying the action (Debnath and Roth, 2021) or mentioning required tools (Anthonio et al., 2022). It may not even mention basic requirements or reasoning steps between instructions. In practice, however, it often depends on our background knowledge whether we can actually draw required inferences (see e.g. Huang and Yang, 2023). Making relevant inferences and background knowledge explicit is particularly important for making instructional texts understandable, so that instructions can be followed by everyone without misunderstanding (Anthonio, 2024).

Deal with Greedy Siblings
(...) They may be quite insistent, so be prepared to stand firm (...)
Original: Keep your valuables in a safe.
Revised: Keep your valuables in a safe *to prevent the possibility of theft*.

Table 1: Example from our data. The explicitation inserted during revision provides additional background.

In recent years, different strands of research addressed the question of how texts can be made easier understandable, including work on text simplification in the news domain (e.g. Srikanth and Li, 2021; Wu et al., 2023), culture-sensitive machine translation of Wikipedia texts (Han et al., 2023), and revisions based on feedback in student assessment (Li et al., 2024). As a starting point for similar work on instructional texts, we analyze cases of textual insertions in how-to guides from wikiHow (Anthonio et al., 2020) and build a sub-corpus of *explicitations*, i.e. cases in which meaning aspects that are presumably inferrable are made explicit (§3). Note the example in Table 1: a reader might be able to infer why valuables should be kept in a safe from context or background knowledge, but the explicit insertion eliminates the need for inference. To get a deeper insight into such cases and how they differ from potential introductions of new information, we perform a number of linguistic analyses of the collected data (§4). We further analyze how well existing models can distinguish between cases of explicitation and other insertions in a corpus with manual annotations to investigate the models' ability to infer implicitly

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 178–197

conveyed meaning (§5).

In sum, our contributions are three-fold: (i) we present an approach for extracting potential cases of explicitation from revisions of instructional texts, (ii) we build a corpus with human judgments on the distinction between implicit vs. new information, and (iii) we assess the performance of computational models regarding the distinction of explicitation and insertion of new information.

2 Related Work

Similar work to ours within existing research include insertions of elaborations in the form of definitions, explanations or clarifications in text simplification (Srikanth and Li, 2021; Wu et al., 2023) as well as insertions of explanations and examples in student essays (Li et al., 2024), aiming at making texts clearer and more specific. Similarly, Han et al. (2023) study pragmatic explicitations in machine translation where implicit background knowledge in the source text is made explicit in the target text to facilitate better cultural understanding (Klaudy, 1993). We adopt this terminology and transfer it into a monolingual setting such that our explicitations can be viewed as a sub-type of elaborations.

With regard to our experimental settings, this work connects to the identification of revision requirements (Roth and Anthonio, 2021) and generation of revisions (Faltings et al., 2021), which have been shown to benefit from so-called edit intentions, i.e. the reasons for which an edit was made. While edit intention classification is thus a closely related task (Ruan et al., 2024; Rajagopal et al., 2022), most previous work in that direction relates to revisions in scientific or educational writing (e.g. Jiang et al., 2022; Zhang et al., 2017).

When one implies meaning without explicitly stating what is meant, the utterer of the message assumes that the audience is able to work out the intended meaning. This relies on the competence to *reason* given the *shared knowledge* and *context* available to the audience (Grice, 1975). In this regard, the issue of understanding implied meaning is closely tied to commonsense reasoning as addressed in tasks such as Natural Language Inference (NLI; Dagan et al., 2006). Yet, pragmatic reasoning and implied entailment have only in recent years become a topic of interest in this domain. For example, Jeretic et al. (2020) introduced ImpPres, a diagnostic NLI dataset for probing whether NLI models recognize pragmatic inferences like scalar implicature and presupposition as entailment.

3 Data & Annotation

As a basis for studying explicitation in instructional text, we first extract potential cases from the revision histories available as part of wikiHowToImprove (Anthonio et al., 2020). This corpus contains multiple versions of the same wikiHow article at different time steps and aligns sentences changed over time, listing a total of 2.7 million revisions.

In the following, we describe the extraction of candidate items (§3.1), our annotation procedure for verifying candidates (§3.2) as well as our final data split, based on the article categories in wikiHow consisting of 19 high-level topics (§3.3).

3.1 Extracting candidate items

We extract revision pairs where text was inserted into the original version, presumably to facilitate better text comprehension. As candidate items, we consider sentence revisions including insertions of at least three consecutive words, allowing additional minor changes like deletion or replacement of one token. We exclude cases where the insertion is longer than the original sentence, constitutes a new independent clause, or is directly drawn from the original context. As additional filters, we test the impact of removing sentences with all caps tokens and insertions of numbers, named entities (NEs), quotation marks, or coordinated lists. Our motivation for using these additional filters is to increase the ratio of potential explicitations, which are our primary focus, by excluding cases that are very likely to introduce genuinely new information.

3.2 Annotation Procedure

We use the crowdsourcing platform Prolific to determine whether a revision presents a case of explicitation or introduces new information. We collect 8 annotations per candidate item. Annotators are provided with instructions and have to pass a qualification test (see Appendix A)[1].

Annotators are presented with two versions of the same text, one containing the original and the other the revised sentence. The target sentence is highlighted in boldface and the insertion in the revised version is marked in blue. As additional context, the shown texts include a maximum of 3 sentences before and after the relevant sentence

[1]Note that we updated the annotation guidelines after a pilot study. The procedure described here corresponds to the second study. Differences are described in Appendix A.

as well as the name of the article from which the text was drawn. Annotators are asked to decide whether the change in the highlighted sentence would affect the understanding of the text for most readers by selecting "No", indicating implicit meaning, or "Yes", indicating the addition of new information. We thus refer to the labels as *Implicit* ("No") and *New Info* ("Yes"). When selecting "No", annotators are further asked to motivate their choice by selecting one or more options from the categories *Context*, *Reasoning*, and *Background Knowledge*, or defining their own category.[2] To ensure high quality annotations, we only accept participants with English as their first language and an approval rate above 97%.

We collect data in two phases: In the **first phase**, we set up a pilot study with 216 revisions, 108 without and 108 with additional filters (cf. §3.1) to test the setup and whether the additional filters are effective in retrieving primarily cases of explicitation. We collect 8 annotations per revision item and present each annotator with 38 samples, including two attention checks in the form of samples annotated during the qualification test, resulting in six annotation batches in total. We find inter-annotator agreement to be low, with a Krippendorff's α of 0.311 for the batch without additional filters and an α of 0.370 with additional filters.[3] While this is to some extent expected as our task is inherently subjective, it highlights the need for a principled strategy that can accommodate both noise via aggregation (see below) as well as valid disagreements in annotation (see §4.2).

To address noise, we employ Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013), a probabilistic model that jointly infers latent true labels and annotator reliability from annotation data. By identifying and downweighting unreliable annotators, it enables more robust consensus labels in settings with heterogeneous annotations compared to majority voting. We thus filtered out annotators with a MACE score below 0.5 and use the latent labels identified by MACE as our gold labels. After filtering, Krippendorff's α for the remaining annotators is 0.461. Table 2 provides basic statistics, showcasing that our additional filters increase the proportion of cases where implicit meaning is made explicit. We thus decided to keep the additional filters for the main (second) phase.

Pilot / Topic	Krippendorff's α	Percentage *Implicit*
Pilot w/o filters	0.483	57%
Pilot w/ filters	0.440	65%
Arts	0.415	69% (↑)
Business	0.472	61%
Cars	0.439	66%
Computers	0.486	58%
Education	0.404	66%
Family	0.413	68%
Food	0.471	65%
Garden	0.494	54%
Health	0.524	62%
Hobbies	0.476	57%
Pets	0.475	57%
Philosophy	0.480	41% (↓)
Relationships	0.415	62%
Sports	0.611 (↑)	59%
Style	0.500	56%
Travel	0.438	65%
Work	0.356 (↓)	62%
Youth	0.450	59%

Table 2: Krippendorff's α and percentage of explicitations by pilot/topic. Topics marked in grey form the OOD test set. Arrows mark the highest/lowest values.

In the **second phase**, we collect annotations for a total of 1,632 revisions for 19 wikiHow topics. For 15 of these topics, we collect annotations for 102 revisions. Again, we collect 8 annotations per sample and this time annotators are presented with 34 items each, such that each of the topics consist of three annotation batches. The four remaining topics did not yield enough items for three annotation batches after candidate extraction and we thus only selected 34 items for each of these topics.[4] Before applying MACE, the topic data yielded a Krippendorff's α of 0.301, after filtering 0.462.

Table 2 shows that inter-annotator agreement is relatively stable across topics, although some variation exists. These differences do not appear to correlate directly with the proportion of implicit cases. For example, *Sports* and *Work* differ by only three percentage points in implicit instances, yet *Sports* achieves the highest agreement while *Work* has the lowest. This suggests that annota-

[2]For a detailed explanation see Appendix A

[3]Results are averaged over all annotation batches.

[4]We also excluded the topic *Holidays and Tradition* as all annotators besides one were flagged as unreliable by MACE.

Split	Pilot, 108 each	+ 102 (or 34*) Items each from 18 topics	Share	Total	Implicit/New
Train	–	Arts, Cars, Computers, Education, Family, Food, Garden, Health, Pets, Philosophy*, Relationships, Sports, Style, Travel*, Work*, Youth	80%	1,150	61%/39%
Dev	w/o filter		10%	247	57%/43%
Test	w/ filter		10%	247	64%/36%
OOD	–	Business, Hobbies		204	59%/41%

Table 3: Structure of the annotated dataset. Topics marked with an asterisk only contain 34 samples instead of 102.

tion difficulty is more likely influenced by topic characteristics or domain familiarity than by the frequency of implicit content alone.

3.3 Data Split

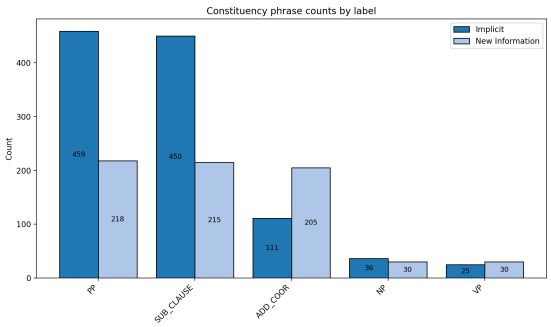
For training, development and testing, we structure the dataset as follows: we select two of the complete topics from the second phase, namely *Business* and *Hobbies*, as 204 out-of-domain (OOD) test articles. From the remaining 16 topics, we sample 80% for our training set and 10% for the development and in-domain test set each. Finally, we add the data from the pilot study to the sets: the batch without additional filters is added to the development set, the one with additional filters to the test set. The final dataset consists of 1,848 items. The split of the dataset and number of samples in each split is illustrated in Table 3. The table shows that the label distribution in our data is fairly similar between splits. Interestingly, there is a lot of variety between the topics. For example, the share of items labeled *New Info* is 58.8% for *Philosophy*, while it is only 31.4% for *Arts*, suggesting that insertions for more complex topics are more likely to be viewed as new information.

4 Analyses

A fundamental assumption underlying our investigation is that explicitations can be distinguished from the insertion of new information. We first analyze the extent to which such a distinction is possible on the basis of linguistic characteristics. To this end, we consider surface features between cases labeled as *Implicit* or *New Info* as well as syntactic and frame-/discourse-semantic structures based on automatic parses (§4.1). On a qualitative level, we examine the extent to which unclear cases in the sense of annotation disagreements can be explained in terms of context, reasoning and background knowledge (§4.2).

Feature	*Implicit*	*New*
Av. Word Length	4.42	4.83
Av. Text Length	25.05	29.14
Lexical Density	0.41	0.49
Freq. Nouns/Adj.	0.23/0.08	0.28/0.11
Freq. Prep./Pron.	0.16/0.11	0.14/0.08

(a) Aggregate linguistic statistics.



(b) Distribution of phrase types in the dataset.

Figure 1: Syntactic analysis of annotated insertions.

4.1 Linguistic Analysis

We analyze the structural linguistic features of the annotated samples to better understand our data and the nature of explicitations. We examine text lengths, the average word lengths, lexical density, the distribution of part-of-speech (POS) categories and phrase types among insertions in revised samples. The results show systematic differences between revisions annotated as *New Info* versus *Implicit*: Specifically, we find revisions labeled *New Info* on average to be longer, exhibiting a higher average word length, and displaying greater lexical density. This is statistically significant using Welch's t-test with p-values below 0.001. In terms of POS distributions, adjectives and nouns occur significantly more frequently in *New Info* revisions, whereas prepositions and pronouns are prevalent in *Implicit* cases, which reflects the typical functions

	Total # (%)	*Implicit* # (%)	*New* # (%)
Revisions	1,848	1,123	725
Parsing units	1,108 (60)	718 (64)	390 (54)
Match RST	826 (45)	523 (47)	303 (42)
Match FN	508 (28)	362 (32)	146 (20)
Match Both	226 (12)	167 (16)	59 (8)

Table 4: Total and relative counts of insertions that match a parsing unit (frame element or discourse unit).

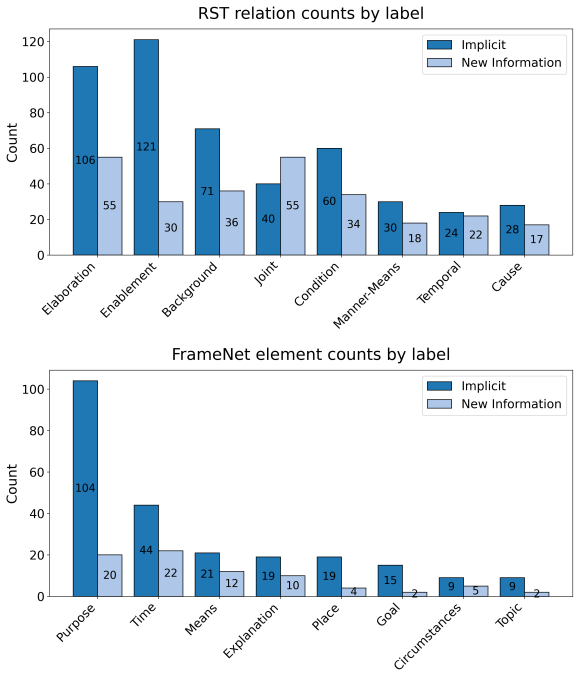


Figure 2: Results of RST and FrameNet parser. Plots illustrate how many revisions correspond to RST relations (top) and FrameNet elements (bottom) respectively.

of these categories: adjectives and nouns tend to contribute substantive semantic content, whereas pronouns primarily refer to previously established entities, and prepositions often serve a more structural role rather than adding novel content. On the phrase level, we find prepositional phrases, subordinate clauses and inserted coordinations to be most frequent, with prepositional and subordinate phrases more often associated with *Implicit* sample and coordination more often associated with *New Info* (see Figure 1).

To further analyze in what way insertions contribute semantically, we utilize two linguistic frameworks designed to represent semantic relations beyond the surface form of a text, namely Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and FrameNet (Baker et al., 1998). While RST is a discourse structure framework for analyzing how different units of a text relate to each other, FrameNet provides the resources for lexical semantics analysis via the identification of so-called frame elements (comparable to semantic roles) tailored to specific frames, i.e. a prototypical situation. We parse the annotated data and extract the samples where the inserted text is mapped to a parsing unit. We use a RST parser by Chistova (2024)[5] and a FrameNet parser by Chanin (2023).[6] Table 4 shows the results, indicating that more than half of the insertions in our data can be mapped to a parsing unit by at least one of the parsers. They further suggest that the frameworks complement each other, as each accounts for different cases with only limited overlap.

In detail, we observe that the insertions in *Implicit* revisions are mapped to the frameworks' parsing units more often than samples labeled *New Info*. In Figure 2, we can see some trends regarding the parsed units and annotation labels. The discourse relation *Enablement* is more frequent for samples labeled as *Implicit*, whereas *Joint* mostly co-occurs with *New Info* following the RST framework. For FrameNet, there is a clear trend for *Implicit* cases being parsed as *Purpose*. RST and FrameNet relations vary across topics and differ in which relation dominates each label. For example, in the topic *Cars*, *Elaboration* occurs only in *Implicit* cases, while *Enablement* is more often labeled as *New Info*, contrary to the overall dataset trend. Similarly for FrameNet *Purpose* is usually implicit overall, but mostly explicit for *Food* and *Pets*. These topic-specific differences might stem from an anticipated difference in domain-familiarity.

4.2 Manual Analysis of Unclear Cases

The annotated data we collected shows a substantial amount of annotator disagreement. While data quality is often equated with high inter-annotator agreement, some scholars argue that annotator disagreement is not noise but signal (Aroyo and Welty, 2013). This stems from the perspective that annotators can have different interpretations of texts and their meaning which can result in valid disagreement. Reasons for such disagreement include vagueness or ambiguity, differences in required knowledge and differing boundaries for categories

[5]https://pypi.org/project/isanlp-rst/
[6]https://github.com/chanind/frame-semantic-transformer

Most Freq. Reason [#]	Example		
	Article Name	Original	Revision
R [12]	(1) Calculate Fat Calories	If the item was homemade, try checking recipe websites for the nutrition information.	If the item was homemade, try checking recipe websites for the nutrition information *of similar recipes*.
C, R [5]	(2) Become a Translator	Becoming a translator takes practice [. . .].	Becoming a translator *of written texts* takes practice [. . .].
C [3]	(3) Drink Creatine	Mix 5 grams with 1 quart (1 liter) of water.	Mix 5 grams with 1 quart (1 liter) of water *if you wish to load*.
BK [2]	(4) Survive a Shark Attack	[This] may result in arterial gas embolism or severe decompression sickness.	[This] may result in arterial gas embolism or severe decompression sickness, *better known as, "the bends"*.

Table 5: Examples of items with substantial disagreement between the annotators from the development set, categorized by most frequent reason for which annotators labeled an item *Implicit*. Acronyms of reasons are **R:** Reasoning, **C**: Context and **BK**: Background Knowledge. Combinations that occurred only once are not listed.

(Weber-Genzel et al., 2024). In order to understand the reasons for disagreements in our data, we examined such cases in our development set manually, taking into account annotators' comments and the categories that they chose when deeming an item *Implicit*. We examined samples from the second study with at least three annotations (after filtering, see §3.2) and where the disagreement ratio was at least 1/3, 25 in total. Table 5 lists the frequency of the reasons that were most commonly selected per item when annotators labeled an item as *Implicit*, together with an example for illustration.

The most frequent reason given for implicit meaning was Reasoning. This intuitively makes sense, as inference relies on many different factors such as background knowledge, cultural and social upbringing, or daily experiences (Grice, 1975), which differ greatly between individuals. As a result, what is perceived as implicit or obvious by one reader may require explicit signaling for another. For instance, people who regularly check nutritional information online may perceive the added text in Example (1) as obvious, while others need the pointer. Some annotators even commented that they themselves perceive the inserted text as not adding new meaning to the item, but recognized that other readers might think differently, thus selecting *New Info*. This kind of variability is well attested in prior work: research in NLI has shown that label variation is common for inference-based tasks (Jiang et al., 2023b; Huang and Yang, 2023). Context and Background Knowledge are selected less frequently as a reason for implicit meaning. However, disagreements on what constitutes common knowledge also leads to unclear cases, as illustrated in Example (4), where the perception of the item hinges solely on whether the inserted text is known and considered common knowledge. Finally, annotators sometimes explicitly noted that they could have "gone either way" on an item, illustrating the fuzzy boundary between implicit meaning and genuinely new information.

These observations suggest that disagreement in our data can serve as a meaningful signal reflecting ambiguity, vagueness or differences in background knowledge, encouraging us to retain such cases.

5 Experiments

To assess the ability of current models to detect implicitely conveyed meaning, we test to what extent existing models can distinguish cases of explicitations from insertions of new information given the data collection described in Section 3. In other words, we evaluate whether models can automatically verify extracted candidates for explicitation.

5.1 Models

We experiment with multiple LLMs in zero-shot and few-shot settings: GPT-5.2 (OpenAI et al., 2024), DeepSeek (DeepSeek-AI, 2024), Mistral-7B-Instruct (Jiang et al., 2023a), Qwen3-4B-Instruct (Team, 2025) and Llama-3-8B-Instruct (Grattafiori et al., 2024). Given the nature of our task, as well as the findings from the analysis of

Model	Accuracy zero	2-shot	4-shot	8-shot	16-shot	32-shot	fine-tuned
Test Set	Majority class accuracy: 0.640						
DeepSeek	0.676	0.688	0.680	0.696	0.696	**0.700**	–
GPT-5.2	**0.773**	0.765	0.753	0.761	0.753	0.725	–
Llama	0.591	0.615	0.587	0.413	0.360	0.024	**0.648**
Mistral	0.656	0.571	0.607	0.457	0.445	0.462	**0.660**
Qwen	0.660	0.603	0.571	0.555	0.490	0.401	**0.664**
OOD Set	Majority class accuracy: 0.588						
DeepSeek	0.667	0.672	**0.706**	0.681	0.681	0.691	–
GPT-5.2	0.745	**0.775**	0.745	0.740	0.740	0.721	–
Llama	0.608	0.588	0.569	**0.613**	0.412	0.005	0.574
Mistral	0.583	0.544	0.525	0.456	0.417	0.392	**0.608**
Qwen	0.647	0.647	0.657	0.652	**0.672**	0.583	0.583

Table 6: Results of LLMs in zero-shot and few-shot settings as well as variants of Llama, Mistral and Qwen that are fine-tuned on the training set. Highest value per line in boldface.

Model	Acc. zero	Acc. ft	F1 zero	F1 ft
Test Set				
BART	0.660	0.660	0.566	**0.664**
ROBERTA	0.623	**0.668**	0.498	**0.673**
DEBERTA	0.587	**0.721**	0.603	**0.730**
Majority	0.640		0.499	
OOD Set				
BART	0.593	**0.642**	0.500	**0.638**
ROBERTA	0.618	**0.706**	0.500	**0.702**
DEBERTA	0.588	**0.701**	0.590	**0.703**
Majority	0.588		0.436	

Table 7: Results of NLI models, off-the-shelf (zero) and fine-tuned (ft).

high-disagreement cases, we also explore three NLI models: a BART large (Lewis et al., 2020) and ROBERTA large model (Liu et al., 2019) fine-tuned on the MNLI dataset[7,8] as well as a DEBERTA large model designed for efficient zero-shot classification on the NLI task (Laurer et al., 2023).[9]

[7]https://huggingface.co/facebook/bart-large-mnli
[8]https://huggingface.co/FacebookAI/roberta-large-mnli
[9]https://huggingface.co/MoritzLaurer/deberta-v3-large-zeroshot-v2.0

5.2 Experimental Settings

We test the LLMs in a zero-shot setting and investigate whether few-shot prompting improves their ability to classify potential instances of implicit meaning. Furthermore, we fine-tune Mistral, Llama and Qwen on our training data. We utilize supervised fine-tuning in a chat-based format. To enable efficient training, we apply LoRA adapters to 8-bit quantized base models and train using early stopping based on validation loss. The NLI models are tested off-the-shelf as well as fine-tuned on the data. While BART and ROBERTA predict the NLI labels *entailment, neutral* and *contradiction*, the DEBERTA model has a binary classification head (*entailment/no entailment*). We interpret *entailment* to indicate implicit meaning and *contradiction/no entailment* as the insertion of new information when models are applied off-the-shelf. As *neutral* does not translate to any of our two labels, we resort to the label with the second highest probability when it is the highest scoring label. For fine-tuning we adapt BART's and ROBERTA's classification head to fit the binary task and tune all parameters of each model. Models are trained with early stopping on the development set, using standard Transformer fine-tuning settings adapted to each architecture.

As evaluation measures, we report the proportion of correctly classified instances (accuracy) as well as weighted F-score, calculated over precision and recall for our two classes.

5.3 Results & Discussion

Tables 6 and 7 summarize the accuracy results for LLMs and NLI models, respectively. The majority baseline already achieves above-chance accuracy on both test splits. Generally, the task appears challenging: among all LLMs, GPT-5.2 achieves the strongest performance, reaching an accuracy of 77.3% in zero-shot setting on the test set and 77.5% in 2-shot setting on the OOD test set, also outperforming the other models in terms of F1 score. Generally, the LLM F1 score results largely reflect the trends in accuracy (see Table 9 in Appendix E).

Across models, few-shot prompting only yields small improvements and sometimes deteriorates performance. For Llama especially, we observe that the model output becomes increasingly verbose with larger input size. Rather than producing the intended labels, the model increasingly generated explanations, imitated contextual content, or produced malformed continuations. This behavior may result from instruction dilution under larger few-shot contexts, causing the model to rely more on patterns and stylistic cues in the prompt rather than consistently following the intended output format. Fine-tuning LLMs leads to the best accuracy scores on the in-distribution test set. On the OOD test set, however, only Mistral benefits from additional fine-tuning.

NLI models perform below or just slightly above the majority baseline when applied off-the-shelf, but ROBERTA and DEBERTA substantially increase performance when fine-tuned, outperforming most LLMs (except GPT5.2 and in some cases DeepSeek). This aligns with prior work showing that fine-tuned classification models often outperform instruction-based models on text classification tasks as the latter are optimized for generative flexibility rather than strict label prediction (Azuma et al., 2025; Zhang et al., 2025). Among the fine-tuned models, DEBERTA achieves the highest performance on the test set while ROBERTA excels on the OOD set.

Despite the overall difficulty of the task, the observed gains from fine-tuning indicate that the data contains informative signals for model training. However, the fact that few-shot prompting, unlike fine-tuning, generally does not outperform zero-shot prompting on the regular test set suggests that a limited number of in-context examples is insufficient for generalization, whereas the task can be learned (at least to some extent) by actual weight adaptation. Furthermore, topic distribution may play a role: We observe that few-shot prompting yields better performance on the OOD test set. This could indicate that classification features vary across topics. At the same time, the reduced topic diversity in this set apparently makes it easier for models to infer relevant patterns from the few-shot examples, even though they stem from different topics (those in the training split).

Tell Your Boyfriend You Need Some Space
You may want to write down these reasons for you to reflect upon later. This will help you form answers to your boyfriend's questions about your decision. *Some common examples for wanting space in a relationship are needing some alone time to decompress after a busy week, wanting to focus on a project, or taking care of private family matters.* [. . .]

Table 8: Example of a universally misclassified item: the insertion in green is classified as *Implicit* by models, but perceived as *New Info* by annotators.

5.4 Error Analysis

We first consider the items misclassified by all models in the zero-shot/off-the-shelf setting as the "most challenging" cases: 9 in-domain and 7 OOD test instances. All these cases are instances of *New Info* and correspond to items where annotators disagreed, indicating inherent ambiguity. Except for DEBERTA, the bias towards predicting *Implicit* can also be seen by all models in general (see Figures 7 and 8 in Appendix F). Few-shot prompting and, in some cases, fine-tuning can reduce this bias, indicating that our training data enables models to resolve some of the most challenging cases. In particular, only four instances remain in the set of items misclassified by all models after fine-tuning, 4 in-domain and zero OOD test instances, but all of them still cases of *New Info*.

Different reasons can be hypothesized for why some items remain challenging even after fine-tuning. In one case the article title seems to provide sufficient context for an insertion to be classified as *Implicit* by the tested models (see Table 8). In contrast, most annotators apparently do not connect the original phrasing to the context given by the title and view it as general, with the insertion adding *New Info* that further restricts the meaning.

Regarding items misclassified by at least one

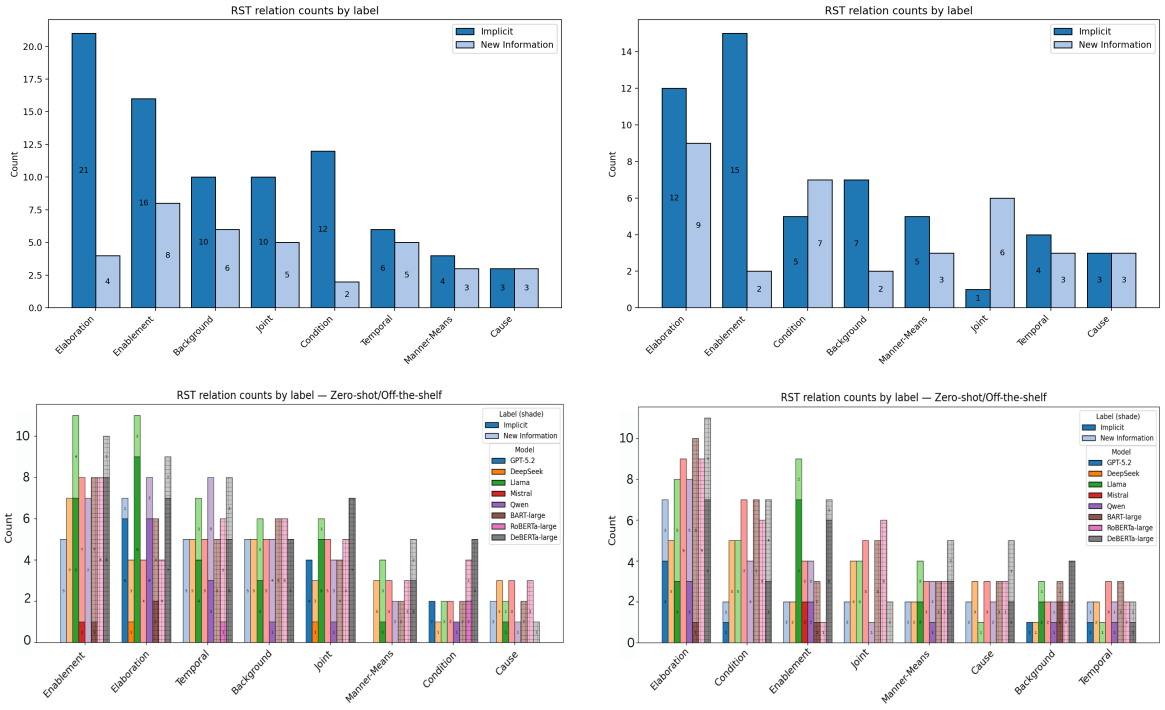


Figure 3: RST relations in the in-domain test (left) and out-of-domain test set (right). The upper row corresponds to all items in the dataset, the bottom row to the items in the set that were misclassified in zero/off-the-shelf setting.

model, we observe that the errors follow the general distribution of phrase types in the corresponding data split (see Figure 10 in Appendix G). This is true for both model types in zero-/off-the-shelf setting and for the LLMs also with increasing few-shots, suggesting that item difficulty does not vary by phrase type. However, model predictions adapt for different phrase structures as training samples are provided. For instance, GPT increasingly misclassifies cases of inserted coordinations as *New Info*, starting in the few-shot setup with 8 samples.

Comparing classification errors across RST units and frame elements yields a different picture, not mirroring the overall distribution in the sets as clearly as the syntactic analysis does. For example, the RST unit TEMPORAL, for instance, is more frequently misclassified on the test set than its general distribution would suggest (see Figure 3). For reference, Figure 11 in Appendix G provides counts for frame elements in the general and misclassified distribution. There is, however, not a clear trend across both test sets, indicating that other aspects like domain might factor into an item's difficulty.

6 Conclusions

We introduce a manually annotated dataset and methodology for studying explicitations in instructional texts. Our analyses show that a distinction between explicitations and insertions of new information is often subjective and closely tied to readers' background knowledge and reasoning abilities. Yet, linguistic analyses revealed systematic differences, indicating that the phenomena are at least in part distinguishable. Our experimental results demonstrate that LLMs achieve strong performance in zero-shot settings, but few-shot prompting does not consistently improve results. While fine-tuning only benefits LLMs in the in-domain setting, NLI models benefit consistently from supervised training and generalize more robustly, indicating that our data provides informative signals. Overall, we view these results and the release of our curated dataset[10] as a crucial step towards the broader goal of automatically enriching instructional texts to improve comprehension and ensure that procedural guidance is accessible to all users.

7 Limitations

Throughout the paper, we hold the assumption that the revisions that the dataset is based on were made with the aim of facilitating better comprehension of the text. Although we filtered out cases of vandal-

[10]Data and code are available at https://github.com/lurr98/classifying_implicit_meaning.

ism where this premise was clearly violated, at this point we cannot verify that even revisions made with the best intentions indeed fulfill the intended effect (cf. Anthonio et al., 2020, for a more general discussion on this topic). For future work, we aim to investigate this further on the released data, for instance by conducting impact evaluations.

Another limitation of our work is that we only performed experiments on five different LLMs, limiting the extent to which our conclusions can be generalized to other models and architectures.

Finally, another shortcoming of our work is that our data is monolingual, focusing only on English as a widely spoken language.

8 Ethical Considerations

We paid annotators on Prolific 11 GBP per hour in the pilot study and 12 GBP per hour for the topic study, which is above the national minimum wage of the country in which this work was conducted. The payment difference is due to the exchange rate fluctuation between study times. Annotators were compensated for their time, even if they did not pass the qualification test. Furthermore, we manually checked the samples provided for annotation to ensure that annotators were not subjected to inappropriate content.

References

Talita Anthonio. 2024. *Linguistically-informed modelling of potentials for misunderstanding*. Ph.D. thesis, PhD Thesis, Universität Stuttgart.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio, Anna Sauer, and Michael Roth. 2022. Clarifying implicit and underspecified phrases in instructional text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3319–3330, Marseille, France. European Language Resources Association.

Lora Aroyo and Christopher Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard.

Daichi Azuma, René Meléndez, Michal Ptaszynski, Fumito Masui, Lara Aslan, and Juuso Eronen. 2025. Svm, bert, or llm? a comparative study on multilingual instructed deception detection. *AI*, 6(9).

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Robyn Carston. 2002. *Pragmatics and Linguistic Underdeterminacy*, chapter 1. John Wiley & Sons, Ltd.

David Chanin. 2023. Open-source frame semantic parsing. *arXiv preprint arXiv:2303.12788*.

Elena Chistova. 2024. Bilingual rhetorical structure parsing with large parallel annotations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9689–9706, Bangkok, Thailand. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alok Debnath and Michael Roth. 2021. A computational analysis of vagueness in revisions of instructional texts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 30–35, Online. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H. P. Grice. 1975. *Logic and Conversation*, pages 41 – 58. Brill, Leiden, Niederlande.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023b. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Kinga Klaudy. 1993. On explicitation hypothesis. In J. Kohn, K. Klaudy, and 1 others, editors, *Transferre Necesse Est... Current Issues of Translation Theory. In Honour of György Radó on His 80th Birthday*, pages 69–77. Dániel Berzsenyi College, Szombathely.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building Efficient Universal Classifiers with Natural Language Inference. *arXiv preprint*. ArXiv:2312.17543 [cs].

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. 2024. Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 365–380, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Michael Roth and Talita Anthonio. 2021. UnImplicit shared task report: Detecting clarification requirements in instructional text. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 28–32, Online. Association for Computational Linguistics.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Re3: A holistic framework and dataset for modeling collaborative document revision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4635–4655, Bangkok, Thailand. Association for Computational Linguistics.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Junyan Zhang, Yiming Huang, Shuliang Liu, Yubo Gao, and Xuming Hu. 2025. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18980–18989, Suzhou, China. Association for Computational Linguistics.

A Instructions

The following instructions were given to the annotators in the **second annotation phase**:

Annotation Guidelines: Implicit Meaning

The goal of this annotation task is to identify sentence pairs in which one sentence contains **implicit meaning** that is made explicit in the other. More specifically, this means that, even though not all information is stated explicitly in the first sentence, **the overall understanding of the text would not change for most readers** when the additional information is added.

The data presented in this task is drawn from a dataset based on *wikiHow* articles. For each item, annotators are first shown the title of the article from which the texts were taken. Below the title, two nearly identical texts are displayed, each containing a sentence highlighted in bold. In the second text, the bold sentence includes an additional element marked in blue. Annotators should ignore other differences between the sentences.

There are two labels available for this task: `Yes` and `No`. During annotation, select `No` if modifying the bold sentence in the given context would not affect the understanding of the text for most readers, even though the original sentence does not make all information explicit.

Indicators for Implicit Meaning

If any of the following conditions apply, select `No`.

1. Context The added information can be inferred from the surrounding context, including the article title.

---

2. Logical Reasoning The added information represents a logical premise or consequence of the original text.

---

3. Background Knowledge The added information is already anticipated based on general background knowledge. This category may sometimes overlap with *Context*.

Indicators for New Info

If any of the following conditions apply, select `Yes`, as they indicate the presence of **new information** rather than implicit content.

1. Addition changes the core meaning The addition fundamentally changes the meaning of the original sentence.

---

2. Added information is too specific The added information introduces specific entities, concepts or events that a regular reader cannot be expected to know about.

> Please note that, since the data is drawn from a *wikiHow* dataset, the text may occasionally sound ungrammatical or unnatural. Annotators should not let this affect their judgment.

(Here are some examples with the correct label:) *This line was added for the few-shot settings, followed by the examples.*

In the first version of the annotation guidelines (during the **pilot study**), Background Knowledge was not listed as a category, instead the following two categories were listed:

---

3. Expected Information The type of information (e.g. a reason, consequence, location) that was added is usually expected by the reader for the specific verb.

4. Recoverable Instruction The instruction remains interpretable even when the information is added such that the same action would be performed from both instructions.

B Prompts

Figure 4 and 5 show an example of the prompts that were given to the LLMs in our experiments.

C Annotation Interface

Figure 6 displays the annotation interface that was shown to the annotators in the second study. In the pilot study, the prompt question was "Does changing the bold sentence affect you understanding of the text?"

D Model Versions

Regarding the LLMs, we used the following versions:

- GPT5.2
- DeepSeek-V3.2 (Reasoning Model)
- Meta-Llama-3-8B-Instruct
- Mistral-7B-Instruct-v0.3
- Qwen/Qwen3-4B-Instruct-2507

E F1 Score

F1 score results of LLMs in zero-shot and few-shot settings as well as fine-tuned variants of Llama, Mistral and Qwen are displayed in Table 9.

F Confusion matrices

Confusion matrices for the zero-shot and fine-tuned models are displayed in Figure 7 and 8.

G Error Analysis Plots

Plots regarding the error analysis are displayed in Figure 10, 3 and 11.

```
# Annotation Guidelines: Implicit Meaning

The goal of this annotation task is to find sentence pairs where one sentence contains
**implicit meaning** that is made explicit in the other. More specifically, this means
that, even though not everything is stated explicitly in the first sentence,
**the understanding of the text would not change for most readers** when the
information is added.

The data that will be presented to you is from a dataset based on wikiHow articles.
For every item you will first be shown the name of the article from which the texts
were taken. Below that you will find two almost identical texts where one sentence
is highlighted in bold. The bold sentence in the second text contains an additional
element, marked by angle brackets <like this>. Do not worry about other changes
in the sentence.

There are two labels for this task: "Yes" and "No".
During the annotation task, select "No" if you think that changing the bold sentence
in the given context would not affect the understanding of the text for most readers,
even though the first text does not state all information explicitly.

---

### Indicators for Implicit Meaning

If any of the following apply, select "No".

#### **1. Context**
The added information is recoverable from the context (including the article title).

---

#### **2. Logical Reasoning**
The added information is a logical premise or consequence of the given text.

---

#### **3. Background Knowledge**
The information in the added text was already anticipated due to existing background knowledge.

---

### Indicators for New Information

If any of the following apply, select "Yes". These suggest **new information**
rather than implicit content:

#### **1. Addition changes the core meaning**
The addition fundamentally changes the meaning of the original sentence.

---

#### **2. Added information is too specific**
The added information introduces specific entities, concepts or events that most readers
cannot be expected to know about.

---

> Please note that, since the data is taken from a wikiHow dataset, the text might
sound ungrammatical or unnatural at times. Do not let this distract you from the task.
```

Figure 4: Prompt containing annotation guidelines and 2 few-shot examples as presented to the LLMs (Part 1)

```
Here are some examples with the correct label:

Example 1:
Article Name: Train_Your_Cat_Not_to_Climb_on_the_Curtains.txt
First Text:
1. **Spray the cat with a squirt of water.** Most cats dislike water so this is a safe, humane
    deterrent.
2. Repeat as necessary. It will have to be repeated numerous times before the cat decides a spray of
    water is not worth the effort of climbing curtains.
3.

Second Text:
1. **Spray the cat <when he climbs on the curtains> with a squirt of water.** Most cats dislike water
     so this is a safe, humane deterrent.
2. Repeat as necessary. It will have to be repeated numerous times before the cat decides a spray of
    water is not worth the effort of climbing curtains.
3.

Label: No


Example 2:
Article Name: Make_a_Canvas.txt
First Text:
12. Set numerous heavy books on top of the flat board to allow the canvas to press; leave this to set
     overnight.
13. **Outline and cut a sheet of brown wrapping paper about 1/2 inch (1.27 cm) shy of the actual size
     of the mounted canvas**
14. Glue the wrapping paper to the back of the canvas panel.
15. Repeat the steps for pressing by laying the panel on a flat surface, placing a flat board on top
    of the panel, and setting heavy books on top of the flat board.

Second Text:
12. Set numerous heavy books on top of the flat board to allow the canvas to press; leave this to set
     overnight.
13. **Outline and cut a sheet of brown wrapping paper about 1/2 inch (1.27 cm) shy of the actual size
     of the mounted canvas <on each side>.**
14. Glue the wrapping paper to the back of the canvas panel.
15. Repeat the steps for pressing by laying the panel on a flat surface, placing a flat board on top
    of the panel, and setting heavy books on top of the flat board.

Label: Yes


*Article name:* Release_Anger.txt

Read the following text and focus on the **bold sentence**.

First Text:

* Draw what you're thinking. It can be a great outlet for everything you have built up inside
    yourself.
* Writing helps to deal with it. **Tell your diary what happened.** You never know, you might end up
    writing a poem or song!

Now read the modified text:

Second Text:

* Draw what you're thinking. It can be a great outlet for everything you have built up inside
    yourself.
* Writing helps to deal with it. **Tell your diary what happened, <without judging your words>.** You
     never know, you might end up writing a poem, or song!

What would most readers say?
Would altering the bold sentence meaningfully change how they understand the text?
```

Figure 5: Prompt containing annotation guidelines and 2 few-shot examples as presented to the LLMs (Part 2)

Model	zero	few2	few4	few8	few16	few32	ft
				F1			
Test Set							
DeepSeek	0.62	0.632	0.619	0.635	**0.659**	0.641	–
GPT	**0.77**	0.767	0.754	0.764	0.757	0.73	–
Llama	**0.592**	0.5	0.586	0.346	0.191	0.047	0.517
Mistral	0.547	0.569	0.588	0.434	0.427	0.433	**0.638**
Qwen	0.64	0.611	0.568	0.525	0.43	0.273	**0.653**
Maj. Bsl.				0.499			
OOD Set							
DeepSeek	0.611	0.622	**0.674**	0.634	0.637	0.651	–
GPT	0.738	**0.773**	0.743	0.737	0.74	0.719	–
Llama	0.596	0.436	0.47	**0.614**	0.24	0.01	0.429
Mistral	0.45	0.543	0.517	0.416	0.302	0.269	**0.58**
Qwen	0.635	0.65	0.66	0.652	**0.672**	0.553	0.569
Maj. Bsl.				0.436			

Table 9: Results of LLMs in zero-shot and few-shot settings as well as variants of Llama, Mistral and Qwen that are fine-tuned on the training set. Highest value per line in boldface.

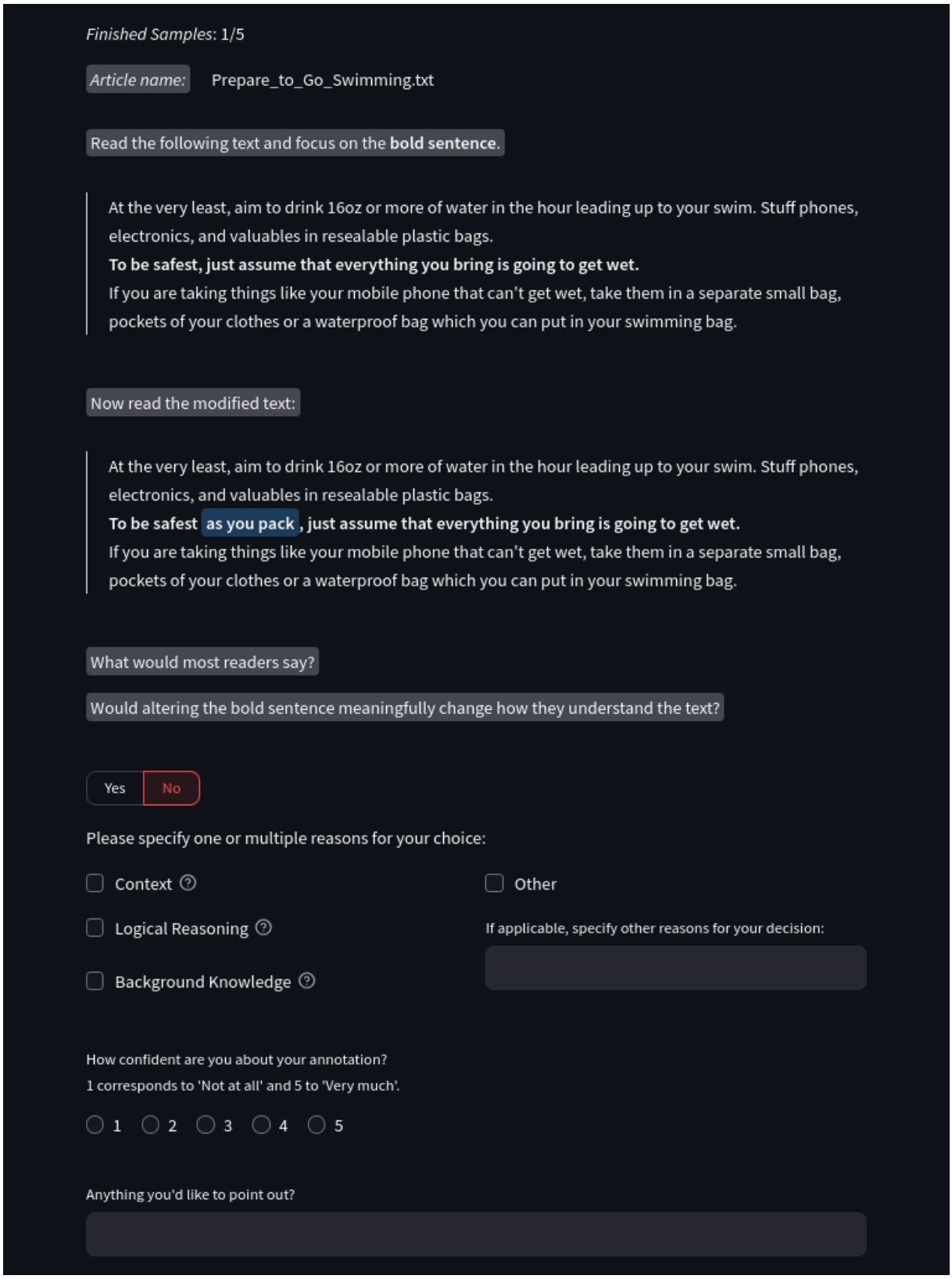


Figure 6: Screenshot of the annotation interface.

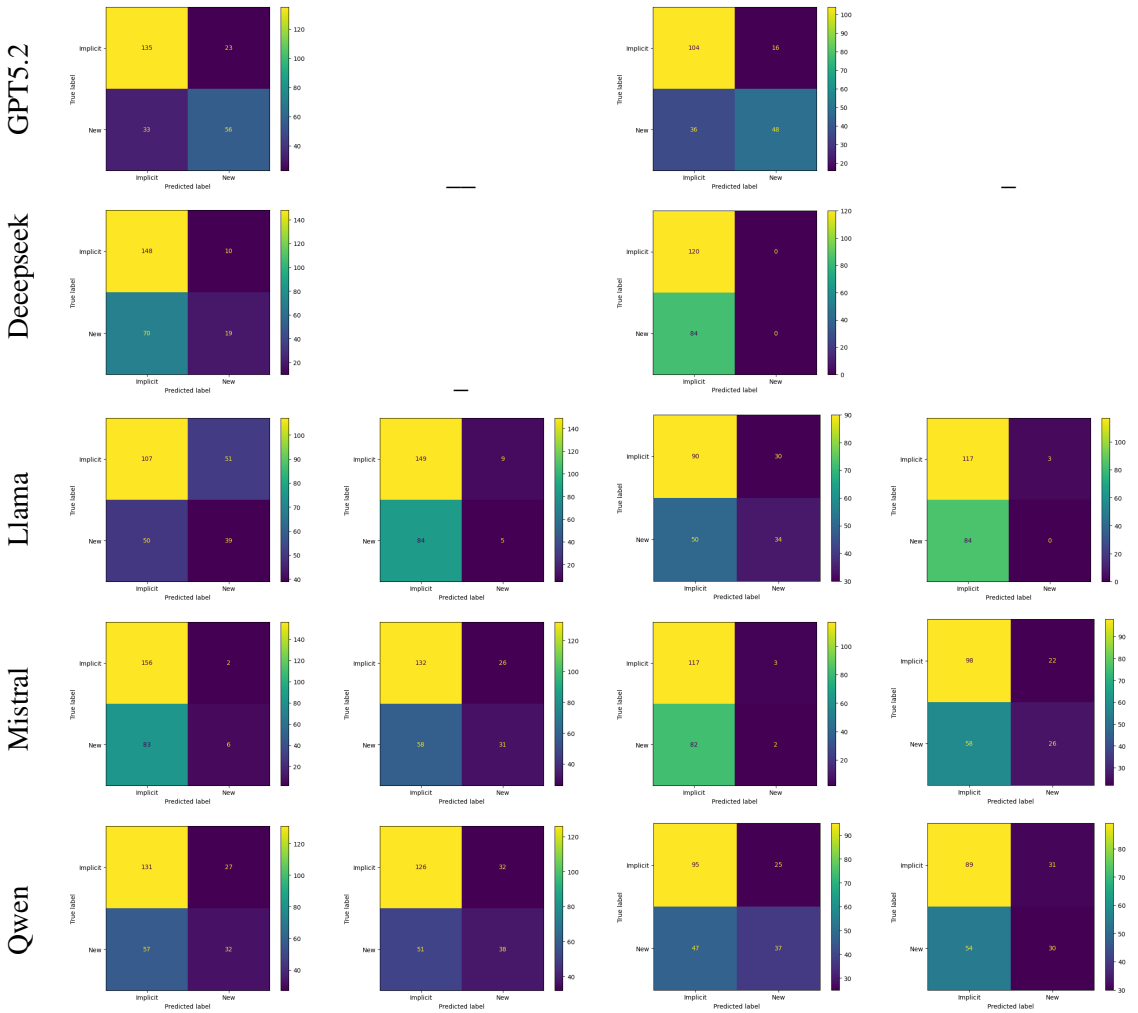


Figure 7: Confusion matrices for the prediction of the LLMs in zero-shot setting on the test set, fine-tuned on the test set, in zero-shot setting on the OOD set and fine-tuned on the OOD set (in this order).

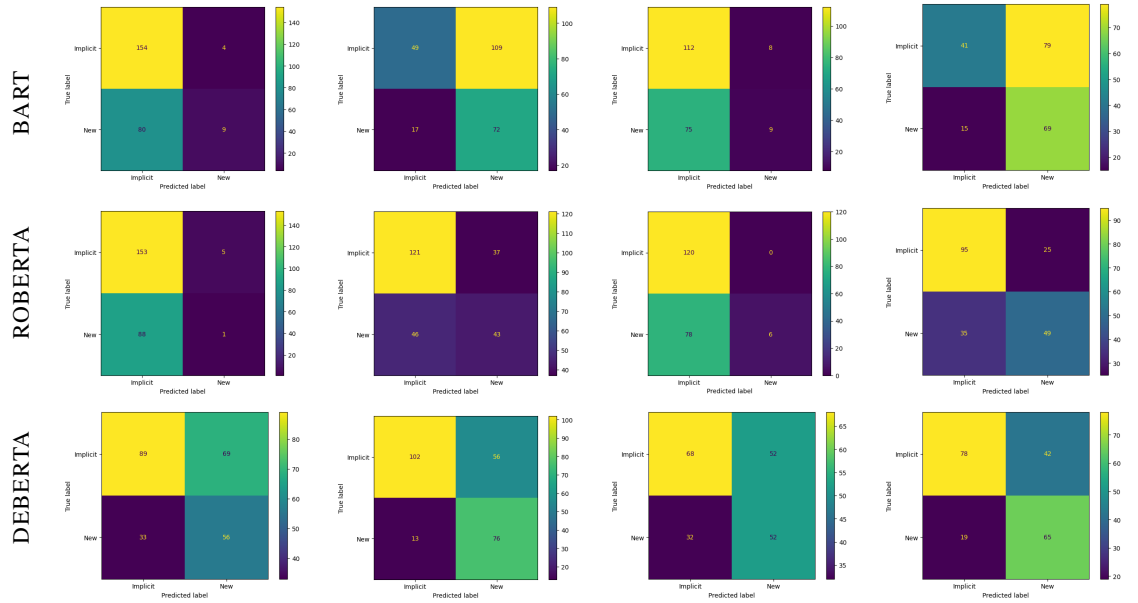


Figure 8: Confusion matrices for the prediction of the NLI models off-the-shelf on the test set, fine-tuned on the test set, off-the-shelf on the OOD set and fine-tuned on the OOD set.

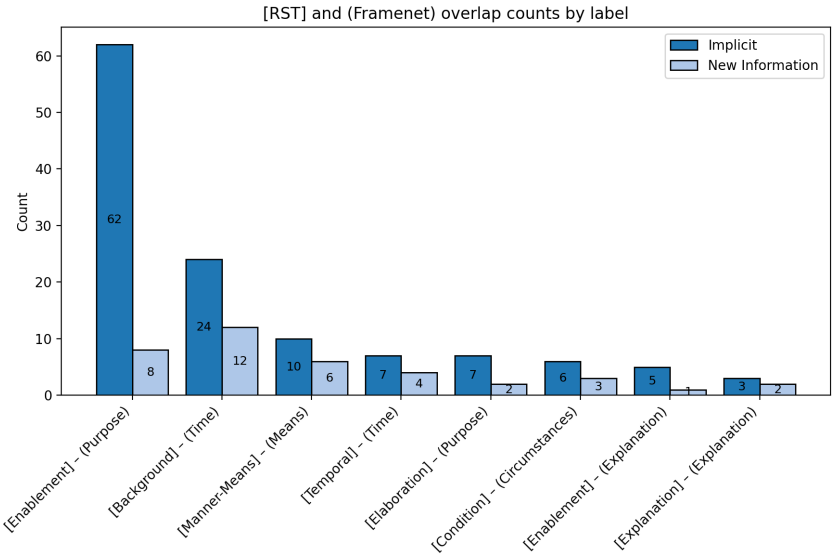


Figure 9: Results from parsing the revisions using an RST and a FrameNet parser, illustrating the overlap, i.e. when a revision yielded a parsed result for both parsers.

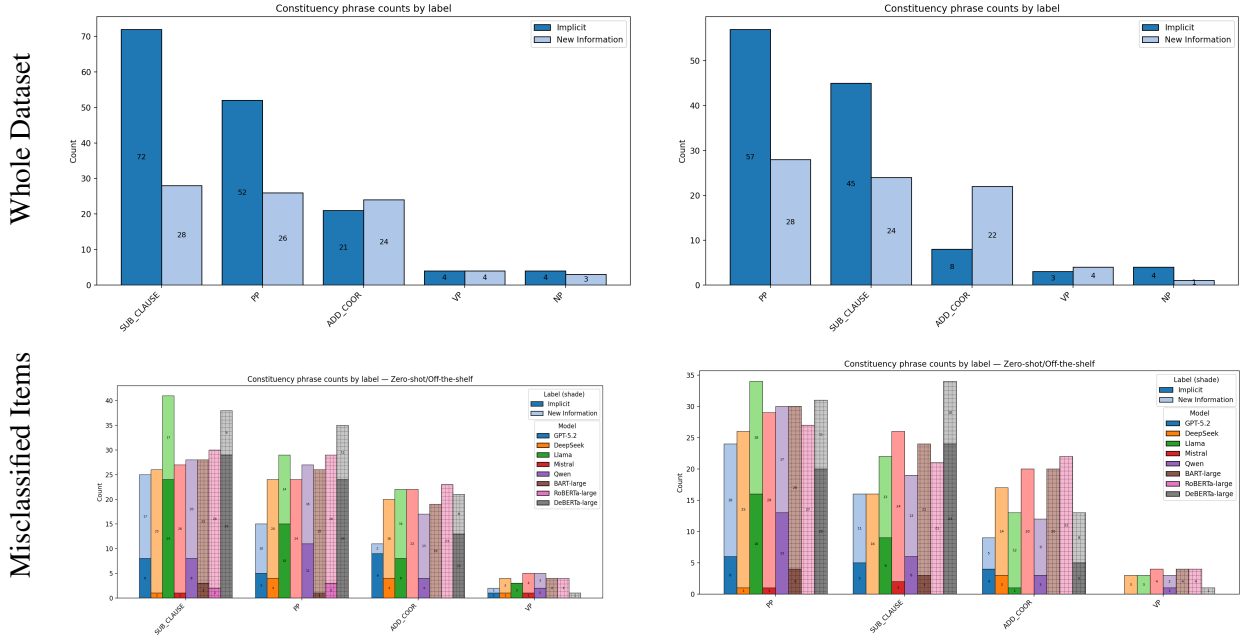


Figure 10: Phrase distribution in the in-domain test (left) and the out-of-domain test set (right). The upper row corresponds to all items in the dataset, the bottom row to the items in the set that were misclassified in zero/off-the-shelf setting.

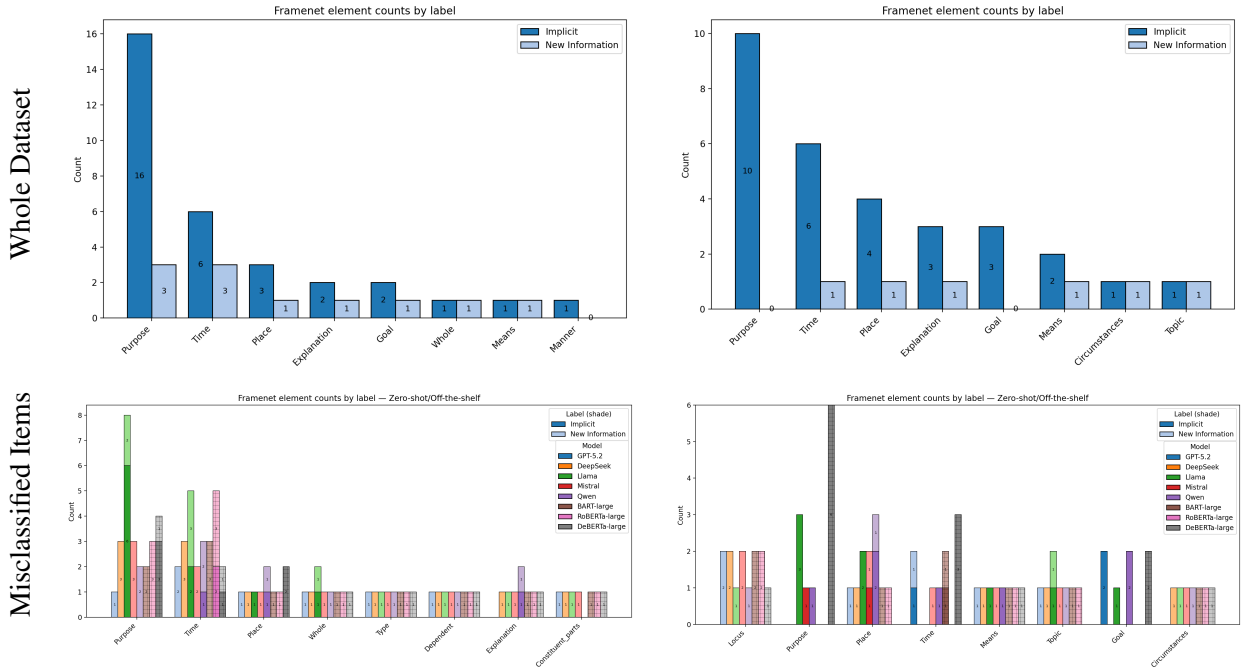


Figure 11: FrameNet elements in the in-domain test (left) and the out-of-domain test set (right). The upper row corresponds to all items in the dataset, the bottom row to the items in the set that were misclassified in zero/off-the-shelf setting.